Eur. Phys. J. B **48**, 279–289 (2005)
DOI: 10.1140/epjb/e2005-00391-0

THE EUROPEAN PHYSICAL JOURNAL B

Extraction of physical laws from joint experimental data

I. Grabec[a]

Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva 6, PP 394, 1001 Ljubljana, Slovenia

Received 5 May 2005 / Received in final form 26 September 2005
Published online 16 December 2005 –

Abstract. The extraction of a physical law $y = y_o(x)$ from joint experimental data about x and y is treated. The joint, the marginal and the conditional probability density functions (PDF) are expressed by given data over an estimator whose kernel is the instrument scattering function. As an optimal estimator of $y_o(x)$ the conditional average is proposed. The analysis of its properties is based upon a new definition of prediction quality. The joint experimental information and the redundancy of joint measurements are expressed by the relative entropy. With the number of experiments the redundancy on average increases, while the experimental information converges to a certain limit value. The difference between this limit value and the experimental information at a finite number of data represents the discrepancy between the experimentally determined and the true properties of the phenomenon. The sum of the discrepancy measure and the redundancy is utilized as a cost function. By its minimum a reasonable number of data for the extraction of the law $y_o(x)$ is specified. The mutual information is defined by the marginal and the conditional PDFs of the variables. The ratio between mutual information and marginal information is used to indicate which variable is the independent one. The properties of the introduced statistics are demonstrated on deterministically and randomly related variables.

PACS. 06.20.Dk Measurement and error theory – 02.50.-r Probability theory, stochastic processes, and statistics – 89.70.+c Information theory and communication theory

1 Introduction

The progress of natural sciences depends on advancement in the fields of experimental techniques and modeling of relations between experimental data in terms of physical laws [1,2]. By utilizing computers a revolution appeared in the acquisition of experimental data while modeling still awaits a corresponding progress. For this purpose the modeling process should be generally described in terms of operations that could be autonomously performed by a computer. A step in this direction was taken recently by a nonparametric statistical modeling of the probability distribution of measured data [3]. The nonparametric modeling requires no a priori assumptions about the probability density function (PDF) of measured data and therefore provides for a fairly general and autonomous experimental modeling of physical laws by a computer [1,4]. Moreover, the inaccuracy of measurement caused by stochastic influences can be properly accounted for in the nonparametric modeling that further leads to the expression of experimental information, redundancy of repeated measurements and model cost function in terms of entropy of information. These variables have already been applied when formulating an optimal nonparametric modeling of PDF, in the most simple case of a one-dimensional variable [3]. However, more frequently than modeling of a PDF the problem is to extract a physical law from joint data about various variables and to analyze its properties. Therefore, the aim of this article is to propose a general statistical approach also to the solution of this problem.

As an optimal statistical estimator of an experimental physical law we propose the conditional average (CA) that is determined by the conditional PDF [1]. This estimator represents a nonparametric regression whose structure is case independent; hence it can be generally programmed and autonomously determined by a computer. Due to these convenient properties, we consider CA as a basis for the autonomous extraction of experimental physical laws in data acquisition systems.

The fundamental steps of the proposed approach to extraction of experimental physical laws from given data are explained in the second section. We first define the estimators of the joint, the marginal and the conditional PDFs and derive from them the conditional average as an optimal estimator of a physical law that is hidden in joint data. In order to estimate the number of data appropriate for the extraction of a physical law, we further introduce the statistics that characterize the information provided by joint measurements. In the third section of the article the properties of the CA estimator and the other introduced statistics are demonstrated on cases of deterministically and randomly related data.

[a] e-mail: igor.grabec@fs.uni-lj.si

2 Statistics of joint measurements

2.1 Uncertainty of experimental observation

Without loss of generality we consider a phenomenon that can be quantitatively characterized by two scalar valued variables x and y comprising a vector $\mathbf{z} = (x, y)$. We further assume that the phenomenon can be experimentally explored by repetition of joint measurements on a two-channel instrument having equal spans $S_x = (-L, L)$, $S_y = (-L, L)$. Their Cartesian product $S_{xy} = S_x \otimes S_y$ determines the joint span. We treat a measurement of a joint datum as a process in which the measured object generates the instrument output $\mathbf{z} = (x, y)$. The basic properties of the instrument and measurement procedure can be characterized by a calibration based on a set of objects $\{\mathbf{w}_{kl} = (u_k, v_l); k = 1, \dots\ l = 1, \dots\}$ that represent joint physical units. Using these units, a scale net can be determined in the joint span S_{xy} of the instrument. In order to simplify the notation, we further omit the indices of units.

A common property of measurements is that the output of the instrument fluctuates even when calibration is repeated [1,2]. We describe this property by the joint PDF $\psi(\mathbf{z}|\mathbf{w})$, which characterizes the scattering of the instrument output at a given joint unit $\mathbf{w}$. For the sake of simplicity, we consider an instrument whose channels can be calibrated mutually independently. In this case the instrument scattering function is expressed by the product of scattering functions corresponding to both channels $\psi(\mathbf{z}|\mathbf{w}) = \psi(x|u)\psi(y|v)$. Their mean values u, v, and standard deviations σ_x, σ_y represent an element of the instrument scale and the scattering of instrument output at the joint calibration. These values can be estimated statistically by the sample mean and variance of both components measured during repeated calibration by a joint unit $\mathbf{w}$. The standard deviation σ characterizes the uncertainty of the measurement procedure performed on a unit [1,2]. We further consider the most frequent case in which the output scattering does not depend on the channel index and the position $\mathbf{w} = (u, v)$ on the joint scale. In this case it can be expressed as a function of the difference $\mathbf{z} - \mathbf{w} = (x - u, y - v)$ and a common standard deviation $\sigma = \sigma_x = \sigma_y$ as $\psi(\mathbf{z}|\mathbf{w}) = \psi(\mathbf{z} - \mathbf{w}, \sigma)$. We consider scattering of instrument output during calibration as a consequence of random disturbances in the measurement system. When these disturbances are caused by contributions from mutually independent sources, the central limit theorem of the probability theory leads us to the Gaussian scattering function $\psi(\mathbf{z} - \mathbf{w}, \sigma) = \mathrm{g}(x - u, \sigma)\mathrm{g}(y - v, \sigma)$, in which the scattering of a single component is determined by:

$$\psi(x|u) \ = \mathrm{g}(x - u, \sigma) \ = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-u)^2}{2\sigma}\right]. \quad (1)$$

2.2 Estimation of probability density functions

Let us consider a single measurement which yields a joint datum $\mathbf{z}_1 = (x_1, y_1)$. We generally consider $\mathbf{z}$ as a random vector variable and assume that this joint datum appears at the outputs of instrument channels, since it is the most probable at a given state $\mathbf{z}$ of the observed phenomenon and the instrument during measurement. Therefore, we utilize the measured datum $\mathbf{z}_1$ as the center of the probability distribution $\psi(\mathbf{z} - \mathbf{z}_1, \sigma) = \psi(x - x_1, \sigma)\psi(y - y_1, \sigma)$ that represents the corresponding state.

Consider next a series of N repeated, equally prepared, and mutually independent measurements which yield the basic data set $\{\mathbf{z}_i; i = 1, \dots, N\}$. In accordance with the above–given interpretation of measured data we adapt to them the distributions $\{\psi(\mathbf{z} - \mathbf{z}_i, \sigma);\ i = 1, \dots, N\}$. We consider measured data as mutually independent, equally weighted samples of the random variable $\mathbf{z}$. Its joint PDF is estimated by the statistical average over distributions $\{\psi(\mathbf{z} - \mathbf{z}_i, \sigma); i = 1, \dots, N\}$ as:

$$f_N(\mathbf{z}) \ = \ \frac{1}{N} \sum_{i=1}^{N} \psi(\mathbf{z} - \mathbf{z}_i, \sigma). \quad (2)$$

This function represents an experimental model of PDF and resembles Parzen's kernel estimator, which is often used in statistical modeling of PDFs [4,5]. However, in Parzen's modeling the kernel width σ plays the role of an optional smoothing parameter whose value should decrease with the number of data N, which is not consistent with the general properties of measurements. In opposition to this, we consider σ as an instrumental parameter that is determined by the inaccuracy of measurement [3,4]. In the majority of experimental observations σ is a constant during measurements, and hence need not be further indicated in the scattering function ψ.

The estimator of PDF given in equation (2) depends on data and function that are completely determined by experimental procedures. Since we do not use any adjustable parameter the estimator equation (2) is nonparametric and hence free of ambiguity that is introduced by an a priori selected form of the estimator in a parametric treatment. The estimated PDF is a basis for our description of the phenomenon under consideration, therefore our approach can be interpreted as an objective statistical one. This interpretation represents an essential advantage with respect to other approaches that are based on various parametric models [4,6,7]. For the same reason we do not specify parametrically the model of the physical law that governs the phenomenon but consider it as a statistic that can be extracted from the estimated PDF by a less ambiguous mathematical treatment. For this purpose we first derive estimators of marginal and conditional PDFs.

The marginal PDF $f(x)$ of a component x is obtained from the joint PDF $f(\mathbf{z}) = f(x, y)$ by integration over the other component, for example:

$$f(x) \ = \int_{S_y} f(x, y)dy. \quad (3)$$

The conditional PDF of the variable y at a given condition x is then defined by the ratio of the joint PDF and the

marginal PDF of the condition:

$$f(y|x) \;=\; \frac{f(x,y)}{f(x)}. \tag{4}$$

Using the experimental model of joint PDF (2) we obtain for the marginal and conditional PDFs the following kernel estimators:

$$f_N(x) \;=\; \frac{1}{N}\sum_{i=1}^{N}\psi(x-x_i,\sigma) \tag{5}$$

$$f_N(y|x) \;=\; \frac{\sum_{i=1}^{N}\psi(x-x_i,\sigma)\psi(y-y_i,\sigma)}{\sum_{j=1}^{N}\psi(x-x_j,\sigma)}. \tag{6}$$

2.3 Estimation of a physical law

It is often observed that the joint PDF resembles a crest along some line $y=\hat{y}(x)$. We consider $\hat{y}(x)$ as an estimator of a hidden physical law $y=y_o(x)$ that provides for a prediction of a value y from the given value x. If we repeat joint measurements, and consider only those that yield the value x, we can generally observe that corresponding values of the variable y are scattered, at least due to the stochastic character of the measurements. As an optimal predictor of the variable y at the given value x, we consider the value $\hat{y}$ that yields the minimum of the mean square prediction error D at a given x:

$$D \;=\; E[(\hat{y}-y)^2|x] \;=\; \min(\hat{y}). \tag{7}$$

The minimum takes place when $\mathrm{dD}/\mathrm{d}\hat{y}=0$. The solution of this equation yields as the optimal predictor $\hat{y}$ the conditional average

$$\hat{y}(x) \;=\; E[y|x] \;=\; \int_{S_y} y\, f(y|x)dy. \tag{8}$$

By using equation (6) for the conditional probability, we obtain for CA the superposition

$$\hat{y}_N(x) \;=\; \frac{\sum_{i=1}^{N} y_i\psi(x-x_i,\sigma)}{\sum_{j=1}^{N}\psi(x-x_j,\sigma)} = \sum_{i=1}^{N} y_i C_i(x). \tag{9}$$

The coefficients

$$C_i(x) \;=\; \frac{\psi(x-x_i,\sigma)}{\sum_{j=1}^{N}\psi(x-x_j,\sigma)} \tag{10}$$

represent a normalized measure of similarity between the given value x and sample values x_i and satisfy the conditions:

$$\sum_{i=1}^{N} C_i(x) = 1\,, \tag{11}$$

$$0 \le C_i(x) \le 1. \tag{12}$$

The more similar given value x is to a datum x_i, the larger the coefficient $C_i(x)$ is and the contribution of the corresponding term $y_iC_i(x)$ to the sum in equation (9). The prediction of the value $\hat{y}_N(x)$, which best corresponds to the given value x, thus resembles the associative recall of memorized items in the brains of intelligent beings, and therefore could be treated as a basis for the development of computerized autonomous modelers of physical laws and related machine intelligence [1].

Predictor $\hat{y}_N(x)$ in equation (9) is completely determined by the set of measured data $\{\mathbf{z}-\mathbf{z}_i;\, i=1,\ldots,N\}$ and the instrument scattering function ψ. The predictor is not based on any a priori assumption about the functional relation between the variables x and y, as is done for example when a physical law is ambiguously described by some regression function in which parameters are adapted to given data [6,7]. Similarly as the estimator of PDF expressed by equation (2), also the conditional average equation (9) can be treated as a nonparametric regression of an objective character. However, it still depends on parameters $\mathbf{z}_i, \sigma$, but these parameters, as well as the form of the function ψ, are totally determined by measurements. They represent a property of the observed phenomenon and not an ambiguously assumed auxiliary of the modeling. Since the form of the CA predictor does not depend on a specific phenomenon under consideration, it could be considered as a generally applicable basis for statistical nonparametric modeling of physical laws in terms of experimental data in an autonomous computer. It is convenient that equation (9) can be simply generalized to a multi-dimensional case by substituting the condition and the estimated variable by the corresponding vectors [1]. Moreover, it is convenient that the ordering into dependent and independent variables is done automatically by a specification of the condition.

In the field of artificial neural networks the kernel estimator of PDF given by equation (2), could be interpreted as a model of Gaussian-radial-basis-function neural network [6,7]. Similarly, the conditional average estimator equation (9), could be considered as a model of a normalized radial–basis–function neural network. However, a significant difference is in the meaning of the kernel width σ: in our approach σ is given by the properties of the instrument utilized in the acquisition of experimental data, while in the field of artificial neural networks it is considered as an optional smoothing parameter that is adapted by the user of the model based upon some optimization procedure. With respect to this difference our approach appears more in tune with experimental exploration and nonparametric statistical modeling of physical laws.

2.3.1 Description of predictor quality

We can interpret a phenomenon which is characterized by the vector $\mathbf{z}=(x,y)$ as a process that maps the variable x to the variable y. When the variables x and y are stochastic, we most generally describe this mapping by the joint PDF $f(x,y)$. Similarly, we can interpret the prediction of the variable $\hat{y}(x)$ from the given value x as a process that runs in parallel with the observed phenomenon. This process is also generally characterized by the PDF $f(x,\hat{y})$,

while the relation between the variables y and $\hat{y}$ is characterized by the PDF $f(y,\hat{y})$. The better the predictor is, the more the distribution $f(y,\hat{y})$ is concentrated along the line $y=\hat{y}(x)$. For a good predictor we generally expect that the prediction error $E_r = y-\hat{y}$ is close to 0. Since both variables are considered as stochastic ones, we expect that the first and second moments of the prediction error $\mathrm{E}[y-\hat{y}]$, $E[(y-\hat{y})^2]$ are small, while for an exact prediction $E[y-\hat{y}]=0$, and $E[(y-\hat{y})^2]=0$. The second moment of the error is equal to $E[(y-\hat{y})^2] = \mathrm{Var}(y) + \mathrm{Var}(\hat{y}) - 2\mathrm{Cov}(y,\hat{y}) + (m_y - m_{\hat{y}})^2$, where $m_y = E[y]$ and $m_{\hat{y}} = E[\hat{y}]$ denote mean values. If the variables y and $\hat{y}$ are statistically independent and have equal mean values, the covariance vanishes: $\mathrm{Cov}(y,\hat{y})=0$, and $m_y - m_{\hat{y}} = 0$, so that $E[(y-\hat{y})^2] = \mathrm{Var}(y) + \mathrm{Var}(\hat{y})$. Based upon this property we introduce a relative statistic called the *predictor quality* with the formula

$$\begin{aligned} Q &= 1 - \frac{E[(y-\hat{y})^2]}{\mathrm{Var}(y)+\mathrm{Var}(\hat{y})} \\ &= \frac{2\mathrm{Cov}(y,\hat{y})}{\mathrm{Var}(y)+\mathrm{Var}(\hat{y})} - \frac{(m_y - m_{\hat{y}})^2}{\mathrm{Var}(y)+\mathrm{Var}(\hat{y})}. \end{aligned} \tag{13}$$

Its value equals 1 for an exact prediction: $\hat{y}=y$, while it equals 0, if the variables y, $\hat{y}$ are statistically independent and have equal mean values. If the mean values differ: $m_y - m_{\hat{y}} \neq 0$, the quality Q can also be negative.

When the predictor is determined by the conditional average (8), we obtain for its mean value

$$\begin{aligned} m_{\hat{y}} = \mathrm{E}[\hat{y}] &= \int \hat{y} f(x)dx = \int\int y f(y|x) f(x) dxdy \\ &= \int\int y f(y,x) dxdy = \mathrm{E}[y] = m_y. \end{aligned} \tag{14}$$

Since in this case $m_y - m_{\hat{y}} = 0$, we further get

$$Q = \frac{2\mathrm{Cov}(y,\hat{y})}{\mathrm{Var}(y)+\mathrm{Var}(\hat{y})}. \tag{15}$$

Similarly we get for the covariance

$$\begin{aligned} \mathrm{Cov}(y,\hat{y}) &= \iint (\hat{y}(x) - m_{\hat{y}})(y-m_y) f(x,y) dxdy \\ &= \iint (\hat{y}(x) - m_{\hat{y}})(y-m_y) f(y|x) dy f(x) dx \\ &= \int (\hat{y}(x) - m_{\hat{y}})^2 f(x) dx = \mathrm{Var}(\hat{y}), \end{aligned} \tag{16}$$

so that the expected quality of the CA predictor is

$$Q = \frac{2\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)+\mathrm{Var}(\hat{y})}. \tag{17}$$

In the case when the relation between both components of the vector $\mathbf{z}$ is determined by some physical law $y_o(x)$, and only the measurement procedure introduces an additive noise ν with zero mean $E[\nu]=0$, and variance $E[\nu^2]=\sigma^2$, we can express the variable y as $y = y_o(x) + \nu$. In this case the following equations: $E[(y-\hat{y})^2] = \sigma^2$, $\mathrm{Var}(y) = \mathrm{Var}(\hat{y}) + \sigma^2$ hold, and we get for the expected predictor quality the expression:

$$Q = \frac{2\mathrm{Var}(\hat{y})}{2\mathrm{Var}(\hat{y}) + \sigma^2}. \tag{18}$$

For $\mathrm{Var}(\hat{y}) \gg \sigma^2/2$ we have $Q \approx 1$, while for $\mathrm{Var}(\hat{y}) \ll \sigma^2/2$ we have $Q \approx 0$. In the last case $\hat{y} \approx$ constant, while y fluctuates around this constant, and consequently the prediction quality is low.

Since generally $\mathrm{Var}(y) \geq \mathrm{Var}(\hat{y})$ and $\mathrm{Var}(\hat{y}) \geq 0$, we obtain from equation (17) the inequality $0 \leq Q \leq 1$. It describes a mean property, which need not be fulfilled exactly if the conditional average is statistically estimated from a finite number of samples N; but we can expect that it holds ever more with an increasing N. However, we can generally expect that with an increasing N, the statistically estimated CA ever better represents the underlying physical law $y = y_o(x)$. However, with an increasing N, the cost of experiments increases, and consequently there generally appears the question: "how to specify a number of samples N that is reasonable for the experimental estimation of a hidden law $y_o(x)$?"

2.4 Experimental information

In order to answer the last question, we proceed with the description of the indeterminacy of the vector variable $\mathbf{z}$ in terms of the entropy of information. Following the definitions given for a scalar random variable in the previous article [3], we first describe the indeterminacy of the component x. For this purpose we introduce a uniform reference PDF $\rho(x) = 1/(2L)$ that hypothetically corresponds to the most indeterminate noninformative observation of variable x; or to equivalently prepared initial states of the instrument before executing the experiments in a series of observations. By using this reference and the marginal PDF $f(x)$, we first define the indeterminacy of a continuous random variable by the negative value of the relative entropy [8,9]

$$H_x = -\int_{S_x} f(x) \log\Big(\frac{f(x)}{\rho(x)}\Big)\, dx. \tag{19}$$

Using the expressions for the reference, instrumental scattering function, and experimentally estimated PDF, we obtain the expressions for the uncertainty H_u of calibration performed on a unit u, the uncertainty H_x of the component x, experimental information I_x provided by N measurements of x, and the redundancy R_x of these measurements as follows [3]:

$$\begin{aligned} H_u &= -\int_{S_x} \psi(x,u) \log(\psi(x,u))\, dx - \log(2L), \\ H_x &= -\int_{S_x} f_N(x) \log(f_N(x))\, dx - \log(2L), \\ I_x(N) &= H_x - H_u, \\ R_x(N) &= \log(N) - I_x(N). \end{aligned} \tag{20}$$

Similar equations are obtained for the component y by substituting $x \to y$.

In order to describe the uncertainty of the random vector $\mathbf{z}$, we utilize the reference PDF that is uniform inside the joint span S_{xy}: $\rho(\mathbf{z}) = \rho(x)\rho(y) = 1/(2L)^2$, and vanishes elsewhere. By analogy with the scalar variable we define the indeterminacy of the random vector $\mathbf{z}$ by the negative value of the relative entropy [8]:

$$H_{xy} = -\int\int_{S_{xy}} f(\mathbf{z}) \log\Big(\frac{f(\mathbf{z})}{\rho(\mathbf{z})}\Big)\, dxdy. \tag{21}$$

In the case of a uniform reference PDF we obtain

$$H_{xy} = -\int\int_{S_{xy}} f(\mathbf{z}) \log(f(\mathbf{z}))\, dxdy - 2\log(2L). \tag{22}$$

With this formula we then express the uncertainty of the joint instrument calibration as

$$H_{\mathrm{w}} = -\int\int_{S_{xy}} \psi(\mathbf{z},\mathbf{w}) \log(\psi(\mathbf{z},\mathbf{w}))\, dxdy - 2\log(2L). \tag{23}$$

For $\sigma \ll L$ we obtain from the Gaussian scattering function $\psi(\mathbf{z},\mathbf{z}_i) = \mathrm{g}(x-x_i,\sigma)\mathrm{g}(y-y_i,\sigma)$ the approximation

$$H_{\mathbf{w}} \approx \log\Big(\frac{\sigma^2}{L^2}\Big) + \log\frac{\pi}{2} + 1. \tag{24}$$

The uncertainty of calibration depends on the ratio between the scattering width 2σ and the instrument span $2L$ in both directions. The number $2\log(\sigma/L)$ determines the lowest possible uncertainty of measurement on the given two-channel instrument, as achieved at its joint calibration.

The indeterminacy of the random vector $\mathbf{z}$, which characterizes the scattering of experimental data, is defined by the estimated joint PDF as

$$H_{xy} = -\int\int_{S_{xy}} f_N(\mathbf{z}) \log(f_N(\mathbf{z}))\, dxdy - 2\log(2L) \tag{25}$$

and is generally greater than the uncertainty of calibration described by $H_{\mathbf{w}}$. Since $H_{\mathbf{w}}$ denotes the lowest possible indeterminacy of observation carried out over a given instrument, we define the joint experimental information I_{xy} about vector $\mathbf{z} = (x, y)$ by the difference

$$\begin{aligned} I_{xy}(N) &= H_{xy} - H_{\mathbf{w}} \\ &= -\int\int f_N(\mathbf{z}) \log(f_N(\mathbf{z}))\, dxdy \\ &\quad + \int\int \psi(\mathbf{z},\mathbf{w}) \log(\psi(\mathbf{z},\mathbf{w}))\, dxdy. \end{aligned} \tag{26}$$

Most properties of the uncertainty and information appertaining to a random vector are similar to those in the case of a scalar variable. For example, the reference density $\rho(\mathbf{z})$ can be arbitrarily selected since it is excluded from the specification of the experimental information [3]. Furthermore, the joint experimental information $I_{xy}(1)$ provided by a single measurement is zero. For a measurement which yields multiple samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$ that are mutually separated by several σ in both directions, the distributions $\psi(\mathbf{z},\mathbf{z}_1) = \mathrm{g}(x-x_i,\sigma)\mathrm{g}(y-y_i,\sigma)$ are non-overlapping and the first integral on the right of equation 26 can be approximated as

$$\begin{aligned} -\frac{1}{N}\sum_{i=1}^{N} \int\int \psi(\mathbf{z},\mathbf{z}_i) \log\Big[\frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{z},\mathbf{z}_i)\Big]\, dxdy \approx \\ \log(N) - \int\int \psi(\mathbf{z},\mathbf{z}_1) \log\psi(\mathbf{z},\mathbf{z}_1)\, dxdy \end{aligned} \tag{27}$$

so that we get $I_{xy}(N) \approx \log(N)$. If the distributions $\psi(\mathbf{z},\mathbf{z}_i)$ are overlapping but not concentrated at a single point, the inequality $0 \le I_{xy}(N) \le \log(N)$ holds generally. Similarly as the entropy of information for a discrete random variable, the experimental information describes how much information is provided by N experiments performed by an instrument that is not infinitely accurate [8]. In accordance with these properties the experimental information describes the complexity of experimental data in units of information entropy, which are here *nats*.

When the distributions $\psi(\mathbf{z},\mathbf{z}_i)$ are non-overlapping, N repeated experiments yield the maximal possible information $\log(N)$. However, with an increasing number N, ever more overlapping of distributions $\psi(\mathbf{z},\mathbf{z}_i)$ takes place, and therefore the experimental information $I_{xy}(N)$ increases more slowly than $\log(N)$. Consequently, the repetition of joint measurements becomes on average ever more redundant with an increasing number N. The difference

$$R_{xy}(N) = \log(N) - I_{xy}(N)\,. \tag{28}$$

thus represents the redundancy of repeated joint measurements in N experiments. Since the overlapping of distributions $\psi(\mathbf{z},\mathbf{z}_i)$ increases with an increasing number of experiments, the experimental information on average tends to a constant value $I_{xy}(\infty)$, and along with this, the redundancy increases with N.

The number

$$K_{xy}(N) = \mathrm{e}^{I_{xy}(N)} \tag{29}$$

describes how many non-overlapping distributions we need to represent the experimental observation. With an increasing N, the number $K_{xy}(N)$ tends to a fixed value $K_{xy}(\infty)$ that can be well estimated already from a finite number of experiments. We could conjecture that $K_{xy}(\infty)$ approximately determines a reasonable number of experiments that provide sufficient data for an acceptable modeling of the joint PDF. However, it is still better to determine such a number from a properly introduced cost function of the experimental observation. With this aim we consider the difference $D_{xy}(N) = I_{xy}(\infty) - I_{xy}(N)$ as the measure of the discrepancy between the experimentally observed and the true properties of the phenomenon. An information cost function is then comprised of the redundancy and the discrepancy measure:

$$C_{xy}(N) = R_{xy}(N) + D_{xy}(N). \tag{30}$$

Since the redundancy on average increases, while the discrepancy measure decreases with the number of measurements N, we expect that the cost function $C_{xy}(N)$ exhibits a minimum at a certain number N_o, which could be considered as an optimal one for the experimental modeling of a phenomenon. From the definition of redundancy and the discrepancy measure we further obtain $C_{xy}(N) = R_{xy}(N)+D_{xy}(N) = \log(N)-2I_{xy}(N)+I_{xy}(\infty)$. Since the last term is a constant for a given phenomenon, it is not essential for the determination of N_o, and can be omitted from the definition of the cost function. This yields a more simple version

$$C_{xy}(N) = \log(N) - 2I_{xy}(N), \tag{31}$$

which is more convenient for application since it does not include the limit value $I_{xy}(\infty)$. In a previous article [3] we have proposed a cost function that is comprised from the redundancy and the information measure of the discrepancy between the hypothetical and experimentally observed PDFs. However, such a definition is less convenient than the present one, although the values of N_o determined from both cost functions do not differ essentially. Numerical investigations also show that the optimal number N_o approximately corresponds to $K_{xy}(\infty) = \mathrm{e}^{I_{xy}(\infty)}$ if the distribution of the data points is approximately uniform.

Although the experimental information of a vector variable and its scalar components exhibits similar properties, their values generally do not coincide since the overlapping of distributions $\psi(\mathbf{z}, \mathbf{z}_i)$ generally differs from that of distributions $\psi(x, x_i)$ or $\psi(y, y_i)$. Therefore, the experimental information provided by joint measurements generally differs from that provided by measurements of single components.

2.5 Mutual information and determination of one variable by the other

In order to describe the information corresponding to the relation between variables x, y we introduce conditional entropy. At a given value x we express the entropy pertaining to the variable y by the conditional PDF as

$$H_{y|x} = -\int_{S_y} f(y|x) \log\Big(\frac{f(y|x)}{\rho(y)}\Big)\, dy. \tag{32}$$

If we express in equation (21) the joint PDF by the conditional one $f(\mathbf{z}) = f(y|x)f(x)$ we obtain the following equation:

$$H_{xy} = \overline{H_{y|x}} + H_x \tag{33}$$

in which $\overline{H_{y|x}}$ denotes the average conditional entropy of information

$$\overline{H_{y|x}} = -\int_{S_x} H_{y|x} f(x)\, dx. \tag{34}$$

When we exchange the meaning of the variables we get

$$H_{xy} = \overline{H_{x|y}} + H_y. \tag{35}$$

Based on these equations and equation (26) we obtain the following relation between the joint and the conditional information

$$\begin{aligned} I_{xy} &= \overline{H_{x|y}} + H_y - H_u - H_v \\ &= \overline{I_{y|x}} + I_x = \overline{I_{x|y}} + I_y \end{aligned} \tag{36}$$

where the conditional information is defined by

$$\overline{I_{x|y}} = \overline{H_{x|y}} - H_u \quad \text{or} \quad \overline{I_{y|x}} = \overline{H_{y|x}} - H_v. \tag{37}$$

When the components of the vector $\mathbf{z}$ are statistically independent, the joint PDF is equal to the product of marginal probabilities and the joint information is given by the sum $I_{xy} = I_x + I_y$, which represents the maximal possible information that could be provided by joint measurements. However, when x and y are not statistically independent, the joint information is less than the maximal possible one: $I_{xy} < I_x + I_y$. The difference

$$I_m = I_x + I_y - I_{xy} = I_x - \overline{I_{x|y}} = I_y - \overline{I_{y|x}}. \tag{38}$$

can be interpreted as the experimental information that a measurement of one variable provides about another one and is consequently called the mutual information [8,10–12]. In accordance with the previous interpretation of the redundancy, it follows from the last two terms in equation (38) that the mutual information also describes how redundant on average is a measurement of the variable y at a given x or vice versa. In accordance with the definition of the redundancy of a certain number N of measurements $R_x(N) = \log(N) - I_x$, we further define also the mutual redundancy of N joint measurements

$$R_m(N) = \log(N) - I_m(N)\,. \tag{39}$$

If we then take into account all the definitions of the redundancies and types of information, we obtain the formula:

$$R_{xy}(N) = R_x(N) + R_y(N) - R_m(N). \tag{40}$$

It should be pointed out that redundancies $R_{xy}(N)$, $R_x(N)$, $R_y(N)$, and $R_m(N)$ generally increase with N, while the corresponding experimental information tends to a fixed value that corresponds to the amount of data needed for presenting related variables.

In order to describe quantitatively how well determined the value of the variable y by the value of x is on average, we propose a *relative measure of determination* by the ratio

$$\overline{D_{y|x}} = \frac{I_m}{I_y} = 1 - \frac{\overline{I_{y|x}}}{I_y}. \tag{41}$$

If $\overline{D_{y|x}} > \overline{D_{x|y}}$, the value of the variable x better determines the value of y than vice versa. In this case the variable x could be considered as more fundamental for the description of the phenomenon, and consequently as an independent one. In the case of functional dependence described by a physical law $y = y_o(x)$, the relative measure

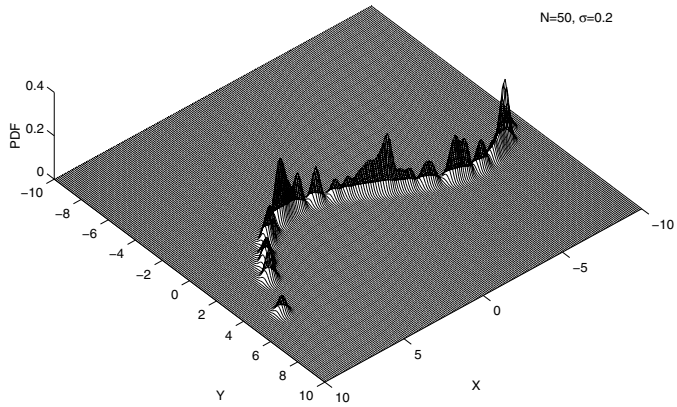

Fig. 1. The joint PDF $f(x, y)$ utilized to demonstrate the properties of the conditional average predictor.

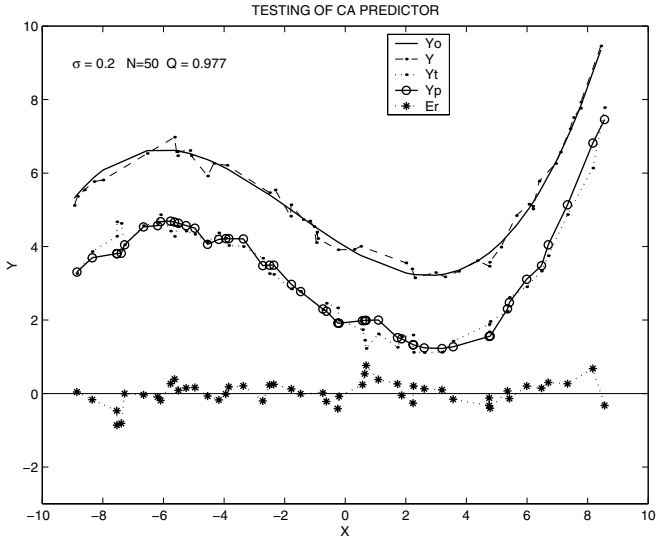

Fig. 2. Testing of CA predictor. Curves representing the underlying law and given data y_o, y – (top), test and predicted data y_t, y_p – (middle), and prediction error $E_r = y_p - y_t$ – (bottom) are displaced in vertical direction for a better visualization.

of determination is $\overline{D_{y|x}} = 1$, while for the statistically independent variables x and y it is $\overline{D_{y|x}} = 0$.

The entropy of information is generally decreased if the distribution of scattered experimental data at a given x is compressed to the estimated physical law $\hat{y}(x)$. The corresponding information gain is in drastic contrast to the information loss that is caused by the noise in a measurement system [13].

3 Illustration of statistics

3.1 Data with a hidden law

The purpose of this section is to demonstrate graphically the basic properties of the statistics introduced above. For this purpose it is most convenient to generate data numerically since in this case the relation between the variables x and y, as well as the properties of the scattering function $\psi(\mathbf{z})$, can be simply set. For our demonstration we arbitrarily selected a third order polynomial law $y_o(x) = [x(x-5)(x+10)]/100$ and the Gaussian scattering function with standard deviation $\sigma = 0.2$. To simulate the basic data set $\{x_i, y_i;\ i = 1, \ldots, N\}$, we first calculated 50 sample values x_i by summing two random terms obtained from a generator with a uniform distribution in the interval $[-8, +8]$ and from a Gaussian generator having the mean value 0 and standard deviation $\sigma = 0.2$. The corresponding sample values y_i were then calculated as a sum of terms obtained from the selected law $y_o(x_i)$ and the same random Gaussian generator with a different seed. The generated data $\{x_i, y_i;\ i = 1, \ldots, 50\}$ were used as centers of scattering function when estimating the joint PDF based on equation (2). An example of such PDF is shown in Figure 1, while the corresponding joint data of the basic set are shown by points in the top curve of Figure 2 together with the underlying law $y_o(x)$.

The conditional average predictor, which corresponds to the presented example, was modeled by inserting data from the basic data set into equation (9). To demonstrate its performance, we additionally generated a test data set by the same procedure as in the case of the basic data set, but with different seeds of all the random generators. Using the values $x_{i,t}$ of the test set, we then predicted the corresponding values $\hat{y}_i$ by the modeled CA predictor. With this procedure we simulated a situation that is normally met when a natural law is modeled and tested based upon experimental data. The test and predicted data are shown by the middle two curves in Figure 2. From both data sets the prediction error $E_r = \hat{y} - y_t$ was calculated that is presented by the bottom curve (..*..) in Figure 2. The curve representing the predicted data (–o–) is smoother than the curve representing the original test data (..·..). This property is a consequence of smoothing caused by estimating the conditional mean value from various data included in the modeled CA predictor. In spite of this smoothing, it is obvious that the characteristic properties of the relation between the variables x and y is approximately extracted from the given data by the CA predictor. This further means that the properties of the hidden law $y = y_o(x)$ can be approximately described in the region where measured data appear based on a finite number of joint samples.

The quality of estimation of the hidden law $y_o(x)$ depends on the values and number N of statistical samples utilized in equation (9) in the modeling of CA and its testing. To demonstrate this property, we repeated the complete procedure three times, using various statistical data sets with increasing N and determined the dependence of predictor quality Q on N. The result is presented in Figure 3. The quality statistically fluctuates with the increasing N, but the fluctuations are ever less pronounced, so that quality determined from different data sets converges to a common limit value at a large N. In our example with $\sigma = 0.2$ the limit value is approximately $Q = 0.98$. With increasing N, the curves corresponding to different data sets join approximately at $N_{\mathrm{CA}} \approx 30$. At a higher N the fluctuations of Q are ever less expressive. We could

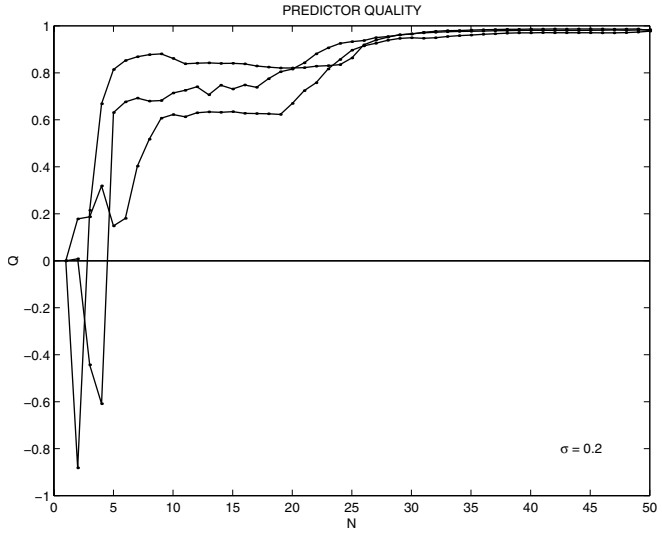

Fig. 3. Dependence of predictor quality Q on number of samples N determined by various statistical data sets.

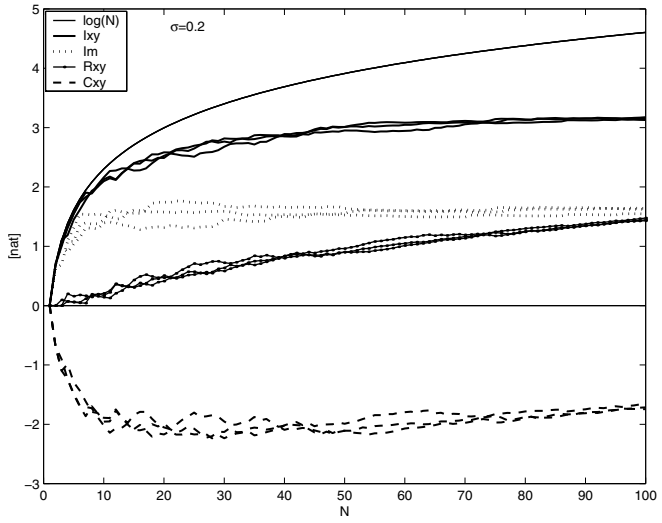

Fig. 4. Dependence of $\log(N)$, experimental information I_{xy}, mutual information I_m, redundancy R_{xy}, and cost function C_{xy} on the number of samples N determined by various statistical data sets.

conjecture that about 30 data values are needed to model the CA predictor in the presented case approximately.

The smaller the scattering width σ is, the higher generally the limit value of the predictor quality is, but on average Q is still less than 1 if $1/\sigma$ and N are finite. This property is in tune with the well–known fact that it is impossible to determine exactly the law $y = y_o(x)$ from joint data that are measured by an instrument which is subject to output scattering due to inherent stochastic disturbances.

The properties of the statistics that are formulated based upon the entropy of information are demonstrated for the case with $\sigma = 0.2$ in Figure 4. It shows the dependence of experimental information I_{xy}, mutual information I_m, redundancy R_{xy}, and cost function C_{xy} on the number of samples N for three different sample

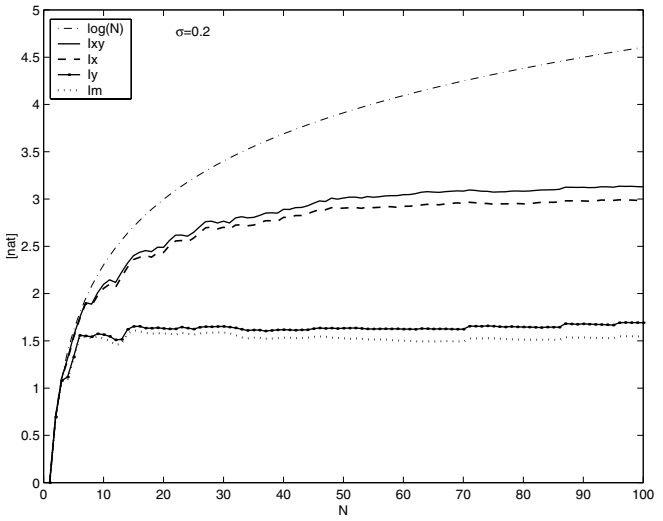

Fig. 5. Dependence of $\log(N)$, experimental information I_{xy}, marginal informations I_x, I_y, and mutual information I_m on the number of samples N.

sets. In the same figure the maximal possible information, which corresponds to the ideal case with no scattering, is also presented by the curve $\log(N)$, since it represents the basis for defining the redundancy. Similarly as in the one-dimensional case [3], the experimental information I_{xy} in the two-dimensional case also converges with increasing N to a fixed value. In the presented case the limit value is $I_{xy}(\infty) \approx 3.2$, which yields the number $K_\infty \approx 25$. This number is approximately equal to the ratio of standard deviation of variable x and the scattering width σ and describes how many uniformly distributed samples are needed to represent the PDF of the data [3]. Due to the convergence of experimental information to a fixed value, the curve $I_{xy}(N)$ starts to deviate from $\log(N)$ with the increasing N. Consequently the redundancy $R_{xy} = \log(N) - I_{xy}(N)$ starts to increase, which further leads to the minimum of the cost function $C_{xy}(N) = \log(N) - 2I_{xy}(N)$. The minimum is not well pronounced due to statistical variations, but it takes place at approximately $N_o \approx 30$. Not surprisingly, the optimal number N_o approximately corresponds to K_∞ and also to N_{CA}.

Similarly as the joint experimental information I_{xy}, the marginal experimental information I_x, I_y also converges to fixed values with increasing N [3]. These statistics are presented in Figure 5 for the same data generator as applied in the case of Figure 4. The sample values of variable x take place in a larger interval than those of variable y. Hence there is less overlapping of scattering functions comprising the marginal PDF of x and consequently I_x is larger than I_y. It is also characteristic that I_{xy} is larger than I_x since the data points in the joint span S_{xy} are more separated than in the marginal span S_x. Since the mutual information I_m is defined as $I_m = I_x + I_y - I_{xy}$, its properties depend on both the marginal and the joint information, and consequently I_m converges more quickly to the limit value than the experimental information I_{xy}.

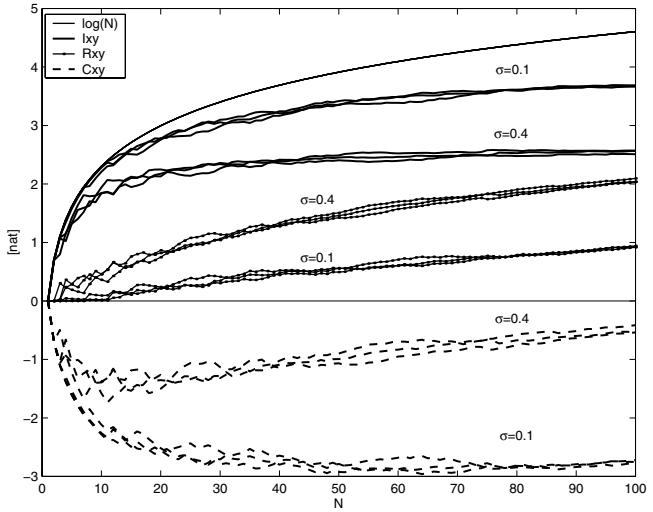

Fig. 6. Dependence of $\log(N)$, experimental information I_{xy}, redundancy R_{xy}, and cost function C_{xy} on the number of samples N determined from various data sets and scattering widths σ.

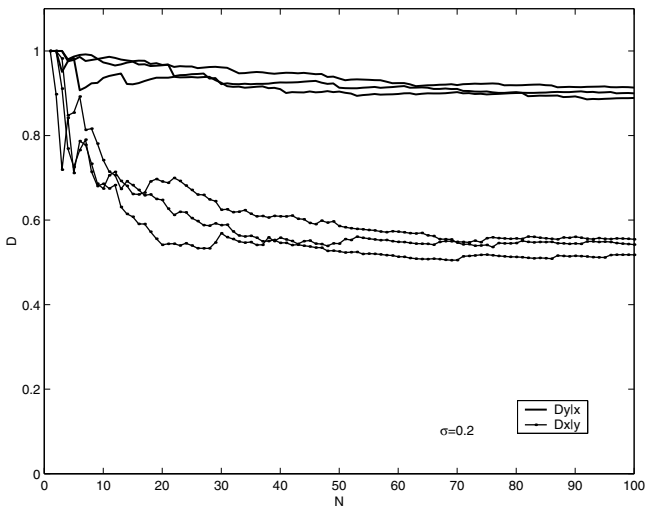

Fig. 7. Dependence of relative measure of determination $\overline{D_{y|x}}$ – (top lines) and $\overline{D_{x|y}}$ – (bottom lines) on the number of samples N determined from various statistical data sets.

To demonstrate the influence of scattering width on the presented statistics the calculations were repeated with $\sigma = 0.1$ and 0.4. The results are presented in Figure 6. For the sake of clear presentation, the curves representing the mutual information I_m are omitted. As could be expected, the limit value of I_{xy} increases with decreasing σ. This property is consistent with the well-known fact that more information can be obtained by experimental observation when using an instrument of higher accuracy that corresponds to a lesser scattering width. In opposition to this, the redundancy of measurement decreases, and along with it, the optimal number N_o increases with the decreasing scattering width.

From the calculated mutual and marginal information, the relative measures of determination $\overline{D_{y|x}}$ and $\overline{D_{x|y}}$ were further determined using various statistical data sets. The results are presented in Figure 7 for the case of scattering width $\sigma = 0.2$. When the number of data N surpasses the interval around the optimal number N_o, statistical variations of $\overline{D_{y|x}}$ and $\overline{D_{x|y}}$ become less pronounced and their values settle close to limit ones. The limit value $\overline{D_{x|y}}$ is essentially lower than $\overline{D_{y|x}}$. This is the consequence of the fact that in our case the variable y is uniquely determined by the underlying law $y_o(x)$ based upon the variable x, but not vice versa. In our case, there are three values of the variable x corresponding to a value of y in a certain interval. Consequently, y is better determined by a given x than vice versa, which further yields $\overline{D_{y|x}} > \overline{D_{x|y}}$. Hence the relative measure of determination indicates that variable x could be considered more fundamental for the description of the relation between the variables x and y.

3.2 Data without a hidden law

To support the last conclusion let us examine an example in which the sample values of the variables x and

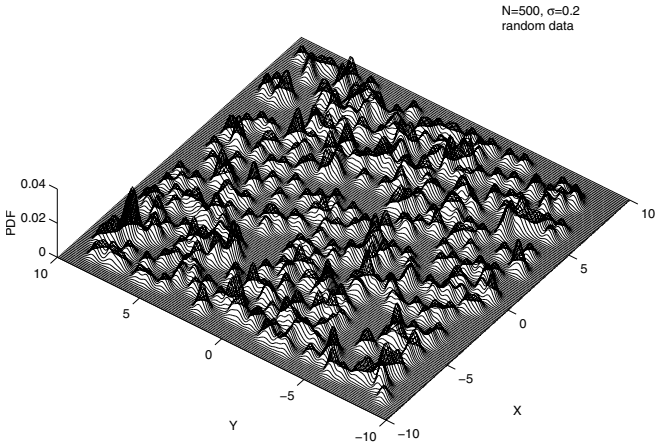

Fig. 8. The joint PDF $f(x, y)$ of $N = 500$ statistically independent random data with $\sigma = 0.2$.

y were calculated by two statistically independent random generators. The corresponding joint PDF is shown in Figure 8, while the properties of the other statistics are demonstrated by Figures 9, 10 and 11.

The properties of the presented statistics could be understood, if the overlapping of scattering functions comprising the estimator of the joint PDF is examined. In the previous case with the underlying law $y_o(x)$, the joint data are distributed along the corresponding line where $-8 \leq x \leq +8$, while in the last case, they take place in the square region $-8 \leq x \leq +8, -8 \leq y \leq +8$. Consequently, the number of samples with non-overlapping scattering functions in the last case is approximately $L/\sigma = 16$ times larger than in the previous case. In the last case we can therefore expect the optimal number of samples in the interval around $N_{ro} \approx 16N_o = 480$. Since in the last case a larger region is covered by the joint PDF, the overlapping of scattering functions is less probable than previously, and therefore, the joint experimental information

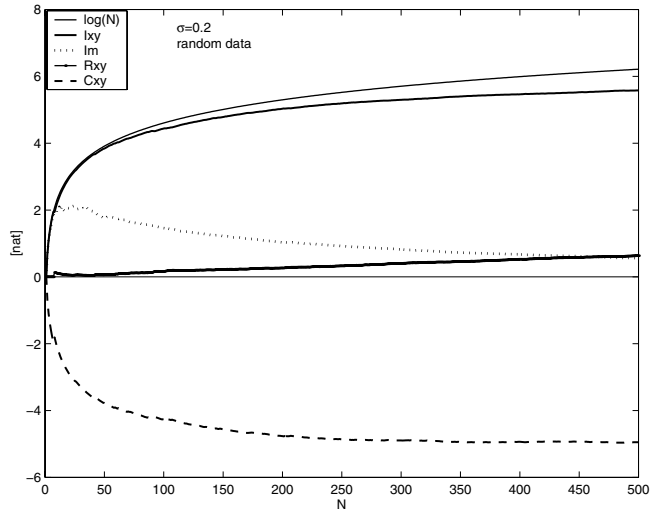

Fig. 9. Dependence of $\log(N)$, experimental information I_{xy}, redundancy R_{xy}, and cost function C_{xy} on the number of samples N in the case of statistically independent random variables x, y.

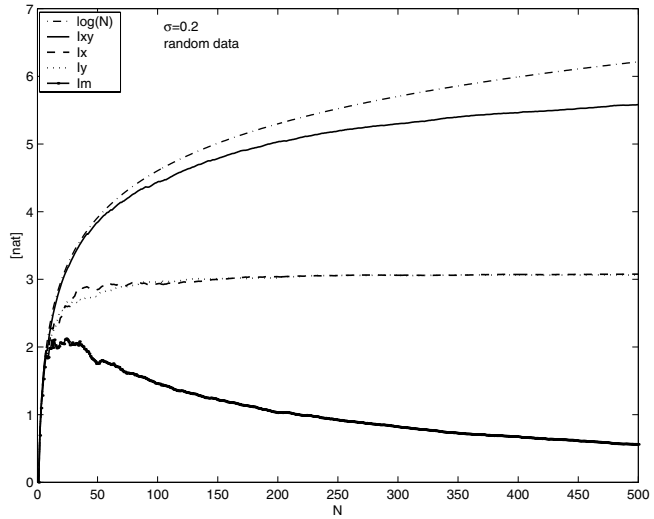

Fig. 10. Dependence of $\log(N)$, experimental information I_{xy}, marginal informations I_x, I_y, and mutual information I_m on the number of samples N in the case of statistically independent random variables x, y.

I_{xy} deviates less quickly from the line $\log(N)$ with the increasing N. Therefore, the redundancy increases less quickly and the minimum of the cost function takes place at a much higher number of $N_{ro} = 480$, which corresponds well to our estimation. Since in the last case the experimental information I_{xy} converges less quickly to the limit value than the marginal information I_x, I_y, the mutual information I_m first increases and later decreases to its limit value. Related to this is the approach of relative measures of determination $\overline{D_{y|x}}, \overline{D_{x|y}}$ to much lower limit values as in the previous case. Since the marginal information I_x, I_y is approximately equal, the curves representing $\overline{D_{y|x}}, \overline{D_{x|y}}$ join with increasing N, and there is

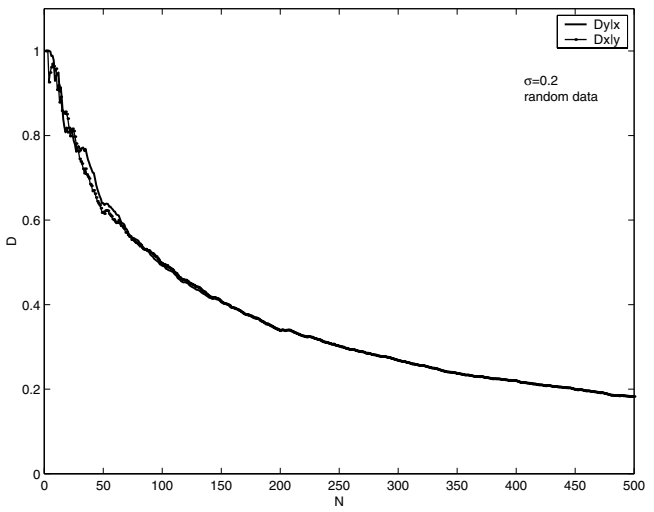

Fig. 11. Dependence of relative measure of determination $\overline{D_{y|x}}$ – (top line) and $\overline{D_{x|y}}$ – (bottom line) on the number of random samples N in the case of statistically independent random data with $\sigma = 0.2$.

no argument to consider any variable as a more fundamental one for the description of the phenomenon under examination. This conclusion is consistent with the fact that the centers of the scattering functions are determined by two statistically independent random generators. However, the limit values of the statistics $\overline{D_{y|x}}, \overline{D_{x|y}}$ are not equal to zero since the region $-8 \leq x \leq +8, -8 \leq y \leq +8$ where the data appear is limited, while the characteristic region $-\sigma \leq x \leq +\sigma, -\sigma \leq y \leq +\sigma$ covered by the joint scattering function does not vanish.

4 Conclusions

Following the procedures proposed in the previous article [3], we have shown how the joint PDF of a vector variable $\mathbf{z} = (x, y)$ can be estimated nonparametrically based upon measured data. For this purpose the inaccuracy of joint measurements was considered by including the scattering function in the estimator. It is essential that the properties of the scattering function need not be a priori specified, but could be determined experimentally based upon calibration procedure. The joint PDF was then transformed into the conditional PDF that provides for an extraction of the law $y_o(x)$ that relates the measured variables x, y. For this purpose the estimation by the conditional average $y_o(x) \approx E[y|x]$ is proposed. The quality of the prediction by the conditional average is described in terms of the estimation error and the variance of the measured data. It is outstanding that the quality exhibits a convergence to some limit value that represents the measure of applicability of the proposed approach. Examination of the quality convergence makes it feasible to estimate an appropriate number of joint data needed for the modeling of the law. It is important that the conditional average makes feasible a nonparametric autonomous extraction of underlying law from the measured data.

Using the joint PDF estimator we have also defined the experimental information, the redundancy of measurement and the cost function of experimental exploration. It is characteristic that experimental information converges with an increasing number of joint samples to a certain limit value which characterizes the number of non-overlapping scattering distributions in the estimator of the joint PDF. The most essential terms of the cost function are the experimental information and the redundancy. During cost minimization the experimental information provides for a proper adaptation of the joint PDF model to the experimental data, while the redundancy prevents an excessive growth of the number of experiments. By the position of the cost function minimum we introduced the optimal number of the data that is needed to represent the phenomenon under exploration. This number roughly corresponds to the ratio between the magnitude of the characteristic region where joint data appear and the magnitude of the characteristic region covered by the joint scattering function. It also corresponds to the appropriate number estimated from the quality of prediction by the conditional average. Based upon the experimental information corresponding to the joint and marginal PDFs, the mutual information has been introduced and further utilized in the definition of the relative measure of determination of one variable by another. This statistic provides an argument for considering one variable as a fundamental one for the description of the phenomenon.

Our method is based upon an experimentally determined scattering function ψ and a measured set of data $\{\mathbf{z}_i;\, i = 1, \ldots, N\}$. However, in the literature there are presented many examples of bare experimental data with no supplementary information about the scattering function ψ and its width σ. In such a case our method cannot be directly applied, since the scattering function can be determined only by a calibration procedure. However, an assumption about the form of scattering function ψ and its width σ can still lead to the application of our method, but in this case it becomes less objective and comparable to other parametric methods [6,7]. Such an assumption also provides for an additional analytical treatment of the properties of PDF estimator. By following the Parzen's approach for the case when $N \to \infty$, we then get to the conclusion that the estimator in equation (2) is a consistent estimator of the hypothetical PDF that is filtered by the kernel function ψ [5,4].

The research was supported by the Ministry of Science and Technology of Slovenia and EU COST.

References

1. I. Grabec, W. Sachse, *Synergetics of Measurement, Prediction and Control* (Springer-Verlag, Berlin, 1997)
2. J.C.G. Lesurf, *Information and Measurement* (Institute of Physics Publishing, Bristol, 2002)
3. I. Grabec, Eur. Phys. J. B **22**, 129 (2001)
4. R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis* (J. Wiley and Sons, New York, 1973), Chap. 4
5. E. Parzen, Ann. Math. Stat. **35**, 1065 (1962)
6. S. Haykin, *Neural Networks* (Prentice Hall International, Inc., Upper Saddle River, New Jersey, 1999)
7. D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003)
8. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1991)
9. A.N. Kolmogorov, IEEE Trans. Inf. Theory **IT-2**, 102 (1956)
10. B.S. Clarke, A.R. Barron, IEEE Trans. Inf. Theory **36**, 453 (1990)
11. D. Haussler, M. Opper, Annals of Statistics **25**, 2451 (1997)
12. D. Haussler, IEEE Trans. Inform. Theory **43**, 1276 (1997)
13. C.E. Shannon, Bell. Syst. Tech. J. **27**, 379 (1948)